Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers

Johan Hattne,[1] Nathaniel Echols,[1] Rosalie Tran,[1] Jan Kern,[1] Richard J. Gildea,[1,10] Aaron S. Brewster,[1] Roberto Alonso-Mori,[2] Carina Glöckner,[3] Julia Hellmich,[3] Hartawan Laksmono,[4] Raymond G. Sierra,[4] Benedikt Lassalle-Kaiser,[1] Alyssa Lampe,[1] Guangye Han,[1] Sheraz Gul,[1] Dörte DiFiore,[3] Despina Milathianaki,[2] Alan R. Fry,[2] Alan Miahnahri,[2] William E. White, [2] Donald W. Schafer,[2] M. Marvin Seibert,[2] Jason E. Koglin,[2] Dimosthenis Sokaras,[5] Tsu-Chien Weng,[5] Jonas Sellberg,[5,6] Matthew J. Latimer,[5] Pieter Glatzel,[7] Petrus H. Zwart,[1] Ralf W. Grosse-Kunstleve,[1] Michael J. Bogan,[2,4] Marc Messerschmidt,[2] Garth J. Williams,[2] Sébastien Boutet,[2] Johannes Messinger,[8] Athina Zouni,[3,9] Junko Yano,[1] Uwe Bergmann,[2] Vittal K. Yachandra,[1] Paul D. Adams,[1] Nicholas K. Sauter[1]

[1]Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.
[2]Linac Coherent Light Source (LCLS), SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.
[3]Max-Volmer-Laboratorium für Biophysikalische Chemie, Technische Universität, D-10623 Berlin, Germany.
[4]Stanford PULSE Institute, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.
[5]Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA.
[6]Department of Physics, AlbaNova, Stockholm University, S-106 91 Stockholm, Sweden.
[7]European Synchrotron Radiation Facility, F-38043 Grenoble Cedex 9, France.
[8]Institutionen för Kemi, Kemiskt Biologiskt Centrum, Umeå Universitet, Umeå, Sweden.
[9]Institut für Biologie, Humboldt Universität zu Berlin, Berlin, Germany.
[10]Current address: Diamond Light Source, Harwell Science and Innovation Campus, Didcot, Oxfordshire OX11 0DE, UK.

Correspondence address: nksauter@lbl.gov (N.K.S.)

Editorial summary:
A computational approach and software tool, *cctbx.xfel*, enables the determination of accurate macromolecular structure factors using a relatively small number of serial femtosecond crystallography diffraction snapshots.

X-ray free-electron laser (XFEL) sources enable the use of crystallography to solve three-dimensional macromolecular structures under native conditions and free from radiation damage. Results to date, however, have been limited by the challenge of deriving accurate Bragg intensities from a heterogeneous population of microcrystals, while at the same time modeling the X-ray spectrum and detector geometry. Here we present a computational approach designed to extract statistically significant high-resolution signals from fewer diffraction measurements.

The ~40 femtosecond-duration XFEL pulse can deliver diffraction information on time scales that outrun radiation damage, allowing macromolecular reaction dynamics to be studied under functional physiological conditions[1-3], while the small beam focus size permits the investigation of extremely small and weakly diffracting microcrystals[4-6]. Unlike single-crystal X-ray diffraction (XRD) experiments performed at conventional synchrotron radiation (SR) sources, XFEL studies destroy the sample with a single pulse, requiring the full data set to be assembled from a series of still diffraction shots of individual microcrystals, a technique known as serial femtosecond crystallography (SFX).

As with conventional crystallography, the objective of SFX is to obtain a complete set of structure factor amplitudes through the measurement of Bragg spot intensities (coherent scattering of X-rays described by Bragg's law) to as high a diffraction angle as possible. The high-resolution signal is ultimately limited by noise, and the background (*e.g.* from solvent) often dominates the diffraction pattern for all but the most intense low-resolution (low angle) Bragg spots[7]. At SR sources, accurate sampling of the diffraction at the limit of detectability is accomplished by optimally modeling the diffraction experiment, including the relationship between real space (the crystal) and reciprocal space (the diffracted X-ray collected on the detector). The most intense Bragg spots are used to deduce the best-fitting lattice model (indexing), which is then used to predict exactly which pixels on each image to examine for Bragg spot integration, even though a signal may not be visually discernable from background. The same fundamental approach is applicable to the analysis of XFEL data. Here, we describe such a data processing approach for XFEL data, which enables weak signals to be measured from many fewer crystal specimens than possible with the previously available program *CrystFEL*[8]. The method has been added to our open source software suite, *cctbx.xfel*[9]. A primer and tutorial may be found at http://cci.lbl.gov/xfel, with code archived at http://cctbx.sf.net.

We tested our method for processing SFX diffraction patterns against data collected at the Coherent X-ray Imaging (CXI) instrument of the Linac Coherent Light Source (LCLS), using the Cornell-SLAC pixel array detector (CSPAD). We derived a structural model for the metalloprotein thermolysin (**Fig. 1a, Supplementary Tables 1 and 2**) that was comparable in quality to structures determined by conventional SR XRD at a similar resolution of 2.1 Å. The electron density of the native calcium and zinc ions (omitted from the phasing model) in the difference map (**Figs. 1b, c**) indicates that the metal positions are determined by the processed data and are not the result of bias from the phasing model. We also reprocessed 1.9 Å-resolution lysozyme data[10] (**Supplementary Tables 1 and 3**) previously processed with the software suite *CrystFEL*[8] to compare the two programs.

We found that *cctbx.xfel* was able to process about twice as many diffraction lattices from individual crystals as previously reported for *CrystFEL*[10] (**Supplementary Table 1**). The indexing algorithm[11] , which identifies unit cell dimensions and crystal orientations, searches for directional vectors that describe the observed rows of Bragg spots, from which three are chosen to form the unit cell. Several factors make this a difficult problem. Firstly, the CSPAD detector consists of 64 pixel array readouts (**Figs. 1d, e**) that are periodically disassembled. Thus the metrology (the relative positions and orientations) of the readouts must be redetermined with sufficient accuracy (**Fig. 2a**), as even small subpixel offsets can diminish the number of images from which lattices can be indexed (**Fig. 2b**). Secondly, the destruction of each crystal after one XFEL shot removes the ability to view the diffracted lattice from various directions, hindering the selection of unit cell basis vectors. To compensate, we supply additional information to the indexing algorithm in the form of a target unit cell, from either an isomorphous crystal form or a preliminary round of indexing. This target unit cell permits us to choose a group of three vectors that best fits the known cell's lengths and angles, thus increasing the number of successfully indexed images. A final factor is the high density of crystals delivered to the X-ray beam, which often produces diffraction patterns containing more than one lattice (**Fig. 1d**). While software exists for modeling multiple lattices in SR diffraction[12, 13], previous XFEL approaches[14] effectively filter these data away, by requiring that 80% of observed spots be covered by a single model. However, we find it straightforward to treat XFEL data with two lattices. The full set of bright candidate Bragg spots is used to derive the first lattice. Candidate spots falling on this lattice are then removed, and the remaining subset is used to find the second lattice, as previously described for SR data[12]. Spot overlaps among multiple lattices were rare, so the minimal inaccuracies in the integrated signal due to overlap were ignored.

The outcome of data integration depends critically on the ability to exactly target the pixels that actually contain signal. A too-inclusive model will capture adjacent pixels that contain only background noise, thus diluting the statistical significance of the measurement. Conversely, overly discriminating models fail to include all of the signal. A crucial first step for data processing, therefore, is to tailor the model to the data at hand. An explanation of why there is a need for new data-modeling algorithms, beyond what is implemented by *CrystFEL*, is presented in the **Supplementary Note.** In short, microscopic "mosaic" domains in the crystal produce Bragg spots shaped like concentric arcs, while the spread of energies in the self-amplified spontaneous emission (SASE) pulse streaks spots radially.

For *cctbx.xfel* we tested two approaches to model the Bragg spots. Although spots vary in size and shape across the lattice (**Fig. 1e**), they tend to be locally similar. This suggests that an empirical approach can be used whereby integration masks are chosen based on the shapes of nearby bright spots. We chose this method—which captures spot shapes of all extremes, both concentric arcs and radial streaks—as the default treatment for data analysis (**Supplementary Tables 1 – 3**). A deeper inspection of the data (**Fig. 2c**) revealed cases where Bragg reflections adjacent to each other nonetheless have very distinct radial widths. These differing widths are explained by the fact that for the full spread of SASE energies to be recorded in the diffraction pattern, Bragg's law demands that the crystal contains microscopic (mosaic) domains with a distribution of either orientations or unit cell dimensions. Wide spots are produced for reflections that satisfy Bragg's law for the full distribution of mosaic domains (given the crystal orientation and range of incident energies), while narrow spots are seen for those reflections that only satisfy the reflecting condition for a subset of domains (**Fig. 2d**). Modeling three parameters (high and low bandpass limits, mosaicity) predicts approximately which pixels to target for signal integration (red dots, **Fig. 2d**). The key benefit of this second, parametric approach is that it roughly accounts for the size and shape differences of adjacent Bragg spots, thus helping the integration mask conform to the actual signal. While the three-parameter model does not give an exact match to the spot shape (**Fig. 2d**), refinement of additional parameters could improve the approach.

We next tested how best to determine the resolution limits of the data set. An important consequence of shot-to-shot variability is that each lattice diffracts to a different limiting angle. Before merging the data into a single set of structure factors, we constructed Wilson plots (integrated Bragg spot intensity vs. diffraction angle bin) in order to determine a separate cutoff angle for each lattice. Once the data had been merged we employed an iterative paired-refinement technique[15] to determine the overall highest resolution shell with a measurable information content (**Fig. 2e**). Remarkably, we found that at the highest resolution proven to contain statistically significant signal (2.1 Å), only 1700 lattices contributed to the thermolysin diffraction data, with an average multiplicity of observation of only 4.5 per structure factor (**Supplementary Table 2**). The size of this selected subset is much smaller than for previous high-resolution XFEL crystallography experiments; past experiments using *CrystFEL* have required >10^4 crystals to obtain reliable structure factors[6, 10, 16]. In cases where only 10^2-10^3 diffracting crystals were available, data merging has only been partially successful[5, 17]. Thus our results with *cctbx.xfel* are encouraging as XFEL progress has been limited by both the difficulty of preparing enough crystal specimens, and the limited data acquisition time at the light source.

In summary, our new developments implemented in *cctbx.xfel* include optimal indexing and retention of data from multiple lattices, separate determination of the resolution cutoff for individual lattices, better descriptions of the Bragg spot shape, and accurate detector geometry to permit well-conforming spot shape models. By carefully discriminating between image pixels known to contain diffraction signal, and the surrounding pixels containing only background noise, we were able to derive accurate structure factors with substantially fewer crystal specimen exposures.

We plan future software developments to further improve the final merged set of structure factors. As illustrated in **Fig. 2d**, a present limitation is that XFEL Bragg diffraction gives only a partial measurement of the structure factor, as the crystal is not fully rotated through the reflecting condition. We intend to implement postrefinement models[18, 19] to allow the correction of intensity measurements to their full-spot equivalent. Such a correction requires a detailed knowledge of the incident spectrum. In **Fig. 2e** the range of X-rays is presented as a top-hat function, but in fact the SASE spectrum is stochastic and finely textured[20]. While the X-ray spectra were not available for the data shown here, single shot measurement of the spectrum is possible[20] and will be incorporated into our method in the future. Taken all together, our method will make it easier for XFEL-based experiments to measure small structure factor differences, such as those from anomalous scattering that will enable the *de novo* determination of macromolecular structures. While SFX is presently a challenging technique, its potential payoff in terms of enabling specialized structural and dynamical studies of macromolecules is enormous.

Accession codes

Protein Data Bank: 4OW3 (structure factors and model for thermolysin); Coherent X-ray Imaging Data Bank: ID23 (raw data streams for thermolysin).

Acknowledgements

This work was supported by US National Institutes of Health (NIH) grants GM095887 and GM102520 and Director, Office of Science, US Department of Energy (DOE) under contract DE-AC02-05CH11231 for data-processing methods (N.K.S.); Director, DOE Office of Science, Office of Basic Energy Sciences (OBES), Division of Chemical Sciences, Geosciences, and Biosciences (CSGB) under contract DE-AC02-05CH11231 (J.Y. and V.K.Y.); NIH grant GM055302 (V.K.Y.); and NIH grant P41GM103393 (U.B.). Sample injection was supported by LCLS (M.J.B., D.W.S.) and the AMOS program, CSGB Division, OBES, DOE (M.J.B.), and through the SLAC National Accelerator Laboratory (SLAC) Laboratory Directed Research and Development Program (M.J.B., H.L.). J.M. was supported by the Artificial Leaf Project Umeå (K&A Wallenberg Foundation), the Solar Fuels Strong Research Environment Umeå (Umeå University), Vetenskapsrådet and Swedish Energy Agency (Energimyndigheten). Experiments were carried out at the LCLS at SLAC, an Office of Science User Facility operated for the DOE by Stanford University. We thank A. Perazzo, M. Dubrovin, I. Ofte, and A. Salnikov (LCLS) for collaboration on data analysis, and C. Kenney (SLAC) for expertise related to the CSPAD detector.

Author Contributions

J. Hattne, J.K., J.Y., U.B., V.K.Y., P.D.A., N.K.S. conceived of the new data processing methods and analyzed the data;
J. Hattne, N.E., R.J.G., A.S.B., R.W.G.-K., P.H.Z., M.M., P.D.A., N.K.S. wrote the data processing software;
U.B., J.Y.,V.K.Y., J.K., R.A.-M.,J.M., A.Z., N.K.S., G.J.W., S.B., A.R.F.,A.M.,D.M., D.W.S., W.E.W., M.J.B. designed the experiment;
R.T., C.G., J.Hellmich, D.D., A.L., G.H., J.K., A.Z. prepared samples;
S.B., J.E.K., M.M., M.M.S., G.J.W. operated the CXI instrument;
M.J.B., H.L., R.G.S., J.K., J.M., B.L.-K., S.G., R.T., C.G., J. Hellmich, J.S., D.W.S., A.M., G.J.W. developed, tested and ran sample delivery system;
R.A.-M., U.B., M.J.B., S.B., N.E., R.J.G., P.G., C.G.,S.G., G.H., J.Hattne., J.Hellmich, J.K., J.E.K., H.L., A.L., B.L.-K., D.M., M.M., J.M.,N.K.S., M.M.S., J.S., R.G.S., D.S., R.T., T.-C.W., G.J.W., V.K.Y., J.Y., A.Z. performed the LCLS experiment;
J. Hattne, N.E., J.K., J.Y., U.B., V.K.Y., P.D.A., N.K.S. wrote the manuscript with input from all authors.

Competing Financial Interests

The authors declare no competing financial interests.

References

1. Neutze, R. *et al. Nature* **406**, 752-757 (2000).
2. Alonso-Mori, R. *et al. Proc. Natl. Acad. Sci. USA* **109**, 19103-19107 (2012).
3. Kern, J. *et al. Science* **340**, 491-495 (2013).
4. Chapman, H.N. *et al. Nature* **470**, 73-77 (2011).
5. Koopmann, R. *et al. Nat. Methods* **9**, 259-262 (2012).
6. Redecke, L. *et al. Science* **339**, 227-230 (2013).
7. Bourenkov, G.P. & Popov, A.N. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 58-64 (2006).
8. White, T.A. *et al. J. Appl. Crystallogr.* **45**, 335-341 (2012).
9. Sauter, N.K. *et al. Acta Crystallogr. D Biol. Crystallogr.* **69**, 1274-1282 (2013).
10. Boutet, S. *et al. Science* **337**, 362-364 (2012).
11. Sauter, N.K., Grosse-Kunstleve, R.W. & Adams, P.D. *J. Appl. Crystallogr.* **37**, 399-409 (2004).
12. Sauter, N.K. & Poon, B.K. *J. Appl. Crystallogr.* **43**, 611-616 (2010).
13. Powell, H.R., Johnson, O. & Leslie, A.G. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1195-1203 (2013).
14. Kirian, R.A. *et al. Acta Crystallogr. A* **67**, 131-140 (2011).
15. Karplus, P.A. & Diederichs, K. *Science* **336**, 1030-1033 (2012).
16. Kirian, R.A. *et al. Opt. Express* **18**, 5713-5723 (2010).
17. Johansson, L.C. *et al. Nat. Methods* **9**, 263-265 (2012).
18. Winkler, F.K., Schutt, C.E. & Harrison, S.C. *Acta Crystallogr. A* **35**, 901-911 (1979).
19. Rossmann, M.G. *et al. J. Appl. Crystallogr.* **12**, 570-581 (1979).
20. Zhu, D. *et al. Appl. Phys. Lett.* **101**, 034103 (2012).

Figures

Figure 1 | Thermolysin structure determination at 2.1 Å resolution. (**a**) $2mF_o–DF_c$ electron density contoured at 1 σ (gray mesh) with water molecules shown as red spheres. (**b**) $mF_o–DF_c$ difference density map contoured at +3 σ (green mesh) and –3 σ (red mesh) showing binding sites for two of the four Ca ions and (**c**) the single Zn ion. (**d**) Detail of two crystal lattices found on the same diffraction image. Modeled spot positions assigned to the different lattices are shown in red and blue, respectively. The sample-detector distance of 135 mm corresponds to a resolution of 2.15 Å at the edges. (**e**) Detail from a different diffraction image. Increasing radial spot elongation is observed with distance from the beam center (blue cross).

Figure 2 | Calibration and validation. (**a**) Aggregate relative positions (top) and rotations (bottom) of 32 pairs of application-specific integrated circuits (ASICs), each pair bump-bonded to a pixel array sensor of the CSPAD detector. The two ASICs on each sensor are manufactured to be aligned along the long axis, separated by a 3.0-pixel gap. These calibration results bear out this expectation within the tolerances shown. (**b**) Impact of positional accuracy on the indexing and integration success rate. Perturbing the ASICs away from their true positions reduces both the total number of indexed images (blue) and the number of images that contain successfully integrated reflections at high (1.8-2.2 Å) resolution (red). Error bars are the standard deviation from five different sets of perturbations drawn from a two-dimensional normal distribution with a standard deviation σ_r. Separate perturbations were drawn for each ASIC. Squares: failure to apply final subpixel corrections from iterative least squares refinement. Circles: failure to apply nearest-whole pixel corrections. (**c**) Detail of four Bragg reflections on a thermolysin diffraction pattern, showing pronounced (seven pixel) radial elongation for the *[27 –34 –7]* reflection and lesser elongation for those nearby. Solution of Bragg's law for each pixel (black arrows) identifies the spread of photon energies that contribute to each reflection. Red disks delineate integration masks from a three-parameter model with wavelength limits λ_{high} = 1.297 (9.556 keV) and λ_{low} = 1.313 (9.443 keV), and full-width mosaic spread δ = 0.174°. (**d**) Reciprocal space diagram indicating how different-shaped reflections arise. Reciprocal lattice points (arcs) all have a constant angular extent δ due to their mosaic spread. Points are in reflecting condition if they are within the zone between the high-energy (red) and low-energy (blue) Ewald spheres. Therefore, a greater fraction of the *[27 –34 –7]* mosaic distribution is within the reflecting condition, leading to a reflection that subtends a greater radial angle $\Delta\theta$. (**e**) Paired refinements of the thermolysin structure. Red and green bars indicate the change in R_{work} and R_{free}, respectively, as higher-resolution data are added to the refinement. Dark and light blue bars show changes to the *R*-factors when the newly added high-resolution structure factors are randomly permuted. The data are interpreted as containing statistically significant signal for the resolution shells where ΔR_{free} is continuously negative, *i.e.* out to 2.1 Å.

Online methods

Sample preparation

Lyophilized thermolysin from *Bacillus stearothermophilus* (Hampton Research) was resuspended in 0.05 M NaOH at a concentration of 25 mg/ml. 300 μl of the protein stock was mixed in a 1:1 ratio with 40% PEG 2000, 100 mM MES pH 6.5, 5 mM $CaCl_2$. Crystallization occurred within minutes. The obtained crystals were transferred into 10% PEG 2000, 100 mM MES pH 6.5, 5 mM $CaCl_2$ (buffer A) and then stepwise into buffer A containing 10, 15, 20 and 30% (w/v) glycerol, respectively. Thermolysin concentration was determined spectrophotometrically using an absorbance value *A*=1.83 (1 mg/ml) at 277 nm[21], and a molecular mass of 34.6 kDa[22]. The final protein concentration of the crystal suspension was found to be 20-24 mg/ml. The average size of the obtained crystals was 2 x 3 x 1 μm^3. As judged by microscope images of various batches the size distribution is very narrow. Assuming an average crystal volume of 6 μm^3, 12 monomers per unit cell and a nominal unit cell volume of 1 x 10^6 $Å^3$, 6 x 10^5 unit cells/crystal gives a concentration of about 3.4 x 10^{10} crystals/ml.

Thermolysin data collection

Diffraction experiments were carried out at the CXI instrument at LCLS[23]. We previously reported the use of a nanoflow liquid injector that markedly reduces the requirements on sample amount[24, 25]. The suspension of thermolysin crystals was injected into the interaction region by this electrospun liquid jet, using a 1 m long silica capillary of 50 μm inner diameter, 150 μm outer diameter, outer diameter tapered at both ends (New Objective) with one end in a pressurized cell outside the vacuum chamber of the CXI instrument, dipping into a vial with 100 μl of the crystal suspension. A potential of +2500V (relative to a counter electrode below the interaction region) was applied to the suspension by means of a bare Pt electrode inside the sample vial. The flow rate was on the order of 0.5 μl/min by applying a backing pressure of 124.1 kPa to the suspension.

The CXI instrument was operated at energies of 9.56 and 9.77 keV (**Supplementary Table 1**), the beam intensity was 6x10^{11} photons/pulse, with a mean pulse duration of 47 fs and a frequency of 120 Hz. The beam was focused to a size of 2.25 μm^2 FWHM at the interaction point. Diffraction was measured utilizing the front CSPAD detector[26] of the CXI instrument. The detector has a pixel size of 110 x 110 μm^2 and a total of 1516 x 1516 pixels.

Resolution of this particular experiment was limited by geometric factors and not the intrinsic strength of the diffraction signal. Several combinations of sample-to-detector distance and incident wavelength were utilized for data collection, but with the most aggressive choice (detector distance = 135 mm, λ = 1.30 Å), geometric limits were 2.15 Å at the detector edge and 1.75 Å in the corner, thus accounting for the falloff in data completeness at high resolution in **Supplementary Table 2**.

Raw data streams have been deposited into the Coherent X-ray Imaging Data Bank[27] (CXIDB; http://cxidb.org) under accession ID 23, along with an exact list of the images that were merged (**Supplementary Tables 1 and 2**) to form the structure factor intensities. A tutorial on accessing information from the raw data files is presented at http://cci.lbl.gov/xfel.

Lysozyme data

To afford a fair comparison between *CrystFEL* and *cctbx.xfel* our only tractable option was to reprocess raw data that had been previously analyzed by the *CrystFEL* software developers. We obtained data from the CXIDB, which archives the raw data streams from the Boutet *et al.* 1.9 Å-resolution structure determination of lysozyme[10] under accession ID 17. To select data for the comparison, we chose only those run numbers (305-327) that yielded the 12,247 images used in the Boutet paper, as documented in a list maintained at the CXIDB Web site (**Supplementary Table 1**). For this particular experiment the CXI instrument was operated at 9.39 keV and the pulse duration was 40 fs. With a detector distance of 93 mm, the geometric limits were 1.74 Å at the detector edge and 1.46 Å in the corner, both well beyond the 1.9 Å resolution limit that we imposed in order to perform a direct comparison with the published results.

Data processing

Data were processed with our package *cctbx.xfel*[9]. After subtraction of a dark-run average image, bright candidate Bragg spots were chosen with the *Spotfinder* component of *cctbx*[28], with settings being adjusted by trial and error specifically for these data; *e.g.,* the minimum spot area was set at 2 square pixels, and the criteria for accepting spots was set to allow spot picking to an outer resolution limit of about 2.5 Å for thermolysin and 1.9 Å for lysozyme. Images were indexed (unit cell dimensions and crystal orientations determined) with the Rossmann DPS algorithm[29, 30] as implemented in our program *LABELIT*[11]. Unit cells dimensions modeled by the indexing algorithm varied from crystal to crystal, with population means and standard deviations for thermolysin reported in **Supplementary Table 1**. A small number of thermolysin lattices (233, ~2%) did not conform to hexagonal Bravais symmetry using our standard criteria[31]; these were removed from further processing and are not included in the reported population. Similarly, 321 non-tetragonal lysozyme lattices were removed (~1%). For previous data analyses with photosystem II[3, 32] we also removed lattices whose unit cell lengths were highly non-isomorphous (differing by >10%) compared to the mean, in order to avoid merging data from non-identical crystal structures[33, 34]. However, for the thermolysin and lysozyme data, none of the unit cell lengths were rejected as outliers.

Improving indexing by using a target unit cell

As stated in the main text, the destruction of each crystal after one XFEL shot makes indexing difficult. Accuracy is much greater at SR sources, where it is possible to mount the crystal on a goniometer and view the diffracted lattice from two different crystal orientations approximately 90° apart[11]. In contrast, the liquid jet method delivers samples in random, unknown, orientations. Furthermore, the XFEL diffraction images examined here varied extensively in quality (resolution and number of Bragg spots), with a less successful indexing outcome from poorer images. With degraded data, the DPS algorithm can fail by choosing three candidate unit cell axes that individually appear to describe periodicity in the diffraction pattern, but when combined do not adequately cover the lattice. To avoid this failure mode, we supplied additional information to the indexing algorithm in the form of a target unit cell taken from isomorphous crystal forms (PDB codes 2TLI for thermolysin and 4ET8 for lysozyme). Groups of three candidate axes from the DPS algorithm are evaluated to find the best fit to the known cell lengths and angles. By requiring this approximate similarity, we increased the number of successfully indexed images from about 8000 to about 11600 for thermolysin. A similar approach was used previously by others to identify the lattice within noisy data[35, 36]. We expect that this method will be generally applicable to XFEL data and not limited to cases where an isomorphous crystal form is known. Data can be treated in two passes, first to determine a consensus unit cell from the highest-quality diffraction images where indexing is readily achieved, and secondly to use this consensus cell as a target for indexing the entire data set. In support of this idea, we note that the population standard deviation of the thermolysin unit cell lengths (**Supplementary Table 1**) is quite narrow (0.3-0.4%), and even for previous low-resolution PS II data[3] the standard deviations (0.9-1.9%) were reasonably low.

Relationship between indexing and hit rates

In a previous paper we described the use of *cctbx.xfel* to provide detailed feedback on the diffraction quality within minutes of data acquisition[9]. For this initial analysis, the *Spotfinder* component of *cctbx*[28] is used to classify a diffraction pattern as a "hit" if it contains 16 or more candidate Bragg spots with dark-subtracted peak heights above 450 analog-digital units (on the CSPAD high-gain setting) out to a resolution limit of 4.0 Å. This peak height criterion is chosen by trial and error to best identify Bragg spots for the thermolysin dataset, and the level can easily be changed in a configuration file for other datasets. **Supplementary Figure 1** shows the final outcome: 77% of the initial low-resolution "hits" are successfully integrated and merged into structure factors; with a slightly lower success rate (65%) for hits containing the lowest number of candidate spots. Reasons for the residual failure rate are still to be determined, and will likely vary from case to case in future experiments.

Empirical approach to modeling the spot shape

Bragg spots from both datasets (thermolysin is illustrated in **Fig. 1e**) were observed to vary in size and shape both within a single lattice and also from image to image. Therefore, the previously published *CrystFEL* model that treats spots as uniformly round and equally-sized in reciprocal space[14] was judged to be a poor fit to this data. As described in the **Supplementary Note**, the underlying phenomenon treated by that model (large λ/a ratio, where λ is the wavelength of the incident light, and a is the crystal width) does not apply for high-resolution experiments. In fact, it is not possible to identify a single criterion to describe the spot shape throughout the data sets; some images exhibit concentric arcs consistent with mosaic spread[37] (not shown), while other images contain elongation that is chiefly radial (**Fig. 1e**). We do note however that whatever the behavior, spots tend to be locally similar in size and shape within each lattice (with one exception, see below). This suggests an empirical approach to determining the spot model. First, easily identified high-intensity Bragg spots (using the program *Spotfinder*[28]) are used to index the lattice. Next, at each predicted lattice position on the image, a mask is constructed consisting of a union of the ten nearest spot shapes from the *Spotfinder* set, similar to the approach taken by some SR data reduction programs[38]. This mask determines the set of pixels to be used for signal summation (integration). Taking a union of all nearby spot masks helps to increase the number of pixels assigned to each Bragg spot, to avoid missing pixels that actually contain signal. This is necessary because the predicted spot positions are slightly inaccurate due to the use of a monochromatic model; in fact the incident light has a 0.5-1.0% bandpass[39] (as described in the paragraph immediately below). This simple empirical approach was used to derive all the structure factor measurements in **Supplementary Tables 1-3**.

Parametric approach to modeling the spot shape

Given the theoretical framework of Bragg's law, it is possible to interpret the shape and size of Bragg spots in terms of more fundamental experimental properties including the spectral dispersion, the crystal size, and the internal crystal disorder[40-47]. Thus, while the above empirical approach is adequate for the present, a deeper understanding of XFEL Bragg spot shapes may be possible. In images of both thermolysin (**Figs. 1e, 2c**) and lysozyme we observe radial spot elongation that is most pronounced at higher diffraction angles. This is consistent with the protein crystals acting as spectral analyzers, such that each Bragg reflection disperses the broad bandpass SASE pulse (typically 0.2-0.5% bandpass)[39] over a radial line up to several pixels wide. Furthermore, we observe that reflections adjacent to each other (**Fig. 2c**) can nonetheless have very distinct radial widths. The explanation is rooted in the fact that for a spread of energies to be recorded in the diffraction pattern, Bragg's law demands that the crystal contain microscopic (mosaic) domains with a distribution of either orientations or unit cell dimensions. **Fig. 2d** represents each Bragg spot as a spherical cap in reciprocal space (shown as an arc) representing a spread of orientations, as has been done previously[48]. In our experiment, wide spots are produced for reflections that satisfy Bragg's law for the full distribution of mosaic domains in the crystal (given the crystal orientation and range of incident energies), while narrow spots are seen for those reflections that only satisfy the reflecting condition for a subset of microscopic domains (**Fig. 2d**). By modeling three parameters (high and low bandpass limits, plus mosaicity) we were able to predict approximately which pixels to target for signal integration for each Bragg reflection (red dots, **Fig. 2d**). The key benefit of this approach is that it roughly accounts for the size and shape differences of adjacent Bragg spots, reducing the inclusion of non-signal pixels in the integration mask, and thus helping to extract weak signals. While the three-parameter model in **Fig. 2d** does not give an exact match to the spot shape, we believe that further development will improve the approach. Important additional parameters that could be refined include the spectral shape and unit cell variation, while others such as crystal size and beam divergence are probably negligible for experiments performed at the CXI 1 μm focus.

Signal integration and error estimation

Signal intensity I for each Bragg spot was integrated over a set of pixels determined by empirical mask construction as described above. A surrounding set of pixels, twice the size of the signal set, and separated from it by a guard zone two pixels wide, was designated for measuring the local background. This background set was used to fit a least-squares plane for background subtraction as described[49]. The estimated variance $\sigma^2(I)$ of the signal measurement was based on counting statistics[49], using a rough estimate for the CSPAD high-gain value of 7.5 analog-to-digital units per photon. Integrated intensities were then corrected for polarization[50]. It was realized that the data set contained numerous intensity measurements at large negative multiples of $\sigma(I)$, from which we concluded that Poisson statistics did not adequately model the experimental error. Error estimates from each diffraction pattern were therefore inflated by assuming that negative values of $I/\sigma(I)$ are actually decoy measurements (noise only) with a Gaussian distribution centered at zero and with a standard deviation of 1, thus providing a lower bound on modeling errors. This inflation factor is determined separately for each image, and acts to increase the initially determined errors from counting statistics. Negative I values were then removed from the data set, and data on each image were scaled to a reference data set derived from an isomorphous structure (section immediately below). When later merging multiple measurements of the same Miller index, the error was modeled simply by propagating the per-measurement $\sigma(I)$ values in quadrature. Since the systematic error contributions for XFEL data are not fully understood, no other systematic correction or error normalization was attempted. The error model derived here is believed to be entirely different than that used in *CrystFEL*, therefore the respective $I/\sigma(I)$ values for the two programs in **Supplementary Tables 1-3** cannot be compared.

Scaling

Integrated intensities from separate images were scaled to intensities derived from an isomorphous reference structure (PDB codes 2TLI for thermolysin and 4ET8 for lysozyme); this scaling step helped to account for specimen-to-specimen variation in crystal size and pulse power. For projects where no isomorphous reference structure is available, we propose an iterative procedure wherein the data are merged once without scaling to gain an approximate set of merged intensities, which are then used as the reference for rejecting poorly correlated images in the next round.

Different resolution cutoffs for each lattice

An important consequence of shot-to-shot variability is that each lattice diffracts to a different limiting angle; this can be illustrated even within a single image (**Fig. 1d**) where one lattice (red) extends to higher resolution than a second one (blue). For data reduction, we choose a separate limit for integrating each lattice. Integration relies on having an accurate crystal orientation model, which in turn depends on the set of bright candidate Bragg spots found in our case by the program *Spotfinder*[28]. For example, if *Spotfinder* spots extend only to 4 Å on a particular image, the orientational model is not accurate enough to predict the positions of weak spots at 2.5 Å resolution. We have verified this general result through studies on simulated data (results not shown). A very conservative approach is therefore used for integration: for each image separately the radius of integration is extended slightly past the *Spotfinder* limit, and a Wilson plot is constructed (integrated Bragg spot intensity vs. diffraction angle bin), to identify a resolution limit at which average intensity falls below average noise (based on counting statistics). The radius is increased until such a crossover point is found, at which point it is concluded that either there is no more signal to be found, or the model has diverged from the data. When merging multiple measurements together, it would be counterproductive to include high-resolution integrated measurements from beyond this limit where there is no signal, as this would degrade the overall statistical significance. Allowing separate resolution cutoffs for each image leads to a final merged data set with high multiplicity of observation at low resolution and lower multiplicity at high resolution (**Supplementary Table 2**), yet there is confidence that the highest resolution shell contains real signal.

The quality of the reflections merged in this fashion was assessed by calculating the correlation coefficient of semi-datasets merged from odd- and even-numbered images ($CC_{1/2}$)[15]. We note that our multiplicity statistics (**Supplementary Tables 2 and 3**) differ from previously published high-resolution XFEL analyses[6] that report uniform multiplicity counts over all resolution bins, which is the result of applying a single global resolution limit.

Validation of the resolution cutoff

As the data quality gradually decreases at the highest resolution (**Supplementary Table 2**), it would be advantageous to derive a convenient statistical "rule of thumb" to determine the highest resolution that contains valid, merged structure factors. There must be some reasonable cutoff as the multiplicity of observation and the internal correlation coefficient $CC_{1/2}$ decrease, but it needs to be established which cutoff values should be chosen. To provide an objective criterion, we employed the iterative paired-refinement technique suggested by Karplus & Diederichs[15]. Each iteration compares the result of two atomic structure refinements, the first using data only out to a conservative resolution limit, and the second including reflections in the next, higher-resolution shell. The two models are then evaluated against the smaller, low-resolution set of reflections, and the two reliability factors are computed (R_{work} and R_{free}[51]). As long as R_{free} decreases, the added data contribute useful information to the refinement. An increase in R_{work} but unchanged R_{free} indicates that the model has become less overfit. As a negative control, the model is refined a third time adding the same higher-resolution intensities, but with randomly permuted (incorrect) Miller indices in the shell. Analysis of the thermolysin data starting at 3.0 Å, and progressing in steps of 0.1 Å towards the highest-resolution limit (1.76 Å) shows that there is significant information (*i.e.*, R_{free} decreases) out to at least 2.1 Å (**Fig. 2e**), while randomly permuted Miller indices nearly always increase the *R*-factors, as expected. At the 2.1 Å cutoff, the average observational multiplicity of each structure factor is only 4.5, and the correlation coefficient between semi-datasets is 17.0%.

Relationship between resolution and accurate detector model

The empirical and parametric approaches to constructing Bragg spot profiles as outlined above place very stringent requirements on the geometrical modeling (metrology) of the detector. Many diffraction patterns (**Fig. 1**) exhibit Bragg spots that are only one or two square pixels in area, particularly at low resolution. For spot modeling to work as proposed, therefore, the position of each pixel in space must be known to substantially better accuracy than the pixel dimension, however this is a difficult goal for current XFEL detectors due to their unique construction as a mosaic of pixel array sensors[26, 52]. We took a bootstrapping approach starting with approximately known sensor positions, followed by the use of Bragg observations from the entire data set (either thermolysin or lysozyme), to derive more accurate sensor positions and orientations by iterative non-linear least squares positional refinement (section immediately below). This improved metrology allowed us to model the Bragg spots with an r.m.s. deviation (observed spot position *vs*. modeled position) of 0.65 and 1.00 pixels for thermolysin and lysozyme, respectively. Any well-diffracting set of protein crystals would have sufficed for this procedure; it was not necessary for the unit cell or structure to be known ahead of time.

To assess the general importance of accurate detector metrology we carried out an analysis in which the accurately refined sensor positions were intentionally perturbed (**Fig. 2b**). Indexing success depended weakly on metrology (half of the images could still be indexed with a positional perturbation of 3.5 pixels); but high-resolution integration is strongly dependent, with a 30% loss of high-resolution signal resulting from a perturbation of just a single pixel. This is exactly as expected; our empirically-determined integration masks conform very tightly to the spot shape, therefore for the method to work the positions of individual detector tiles need to be accurately known.

We arrive at the same conclusion, by a different route, if we simply reverse the refinement steps of our detector calibration. This outcome (for the thermolysin data) is also plotted in **Fig. 2b**. Reversing the final step of iterative non-linear least squares positional refinement leaves us with sensor positions 0.55 pixels away from their true positions, with consequent loss in both high-resolution and overall data. Reversing the penultimate step (where we determine the nearest whole-integer pixel positions without any sensor rotations) puts the sensors 1.38 pixels away from true, with a further degradation in the results.

Refinement of the detector geometry model (metrology)

The CSPAD detector utilized at the CXI instrument is laid out in a mosaic arrangement consisting of four groups (quadrants) of eight silicon pixel-array sensors[26]. As the quadrants can be translated on mechanical rails, a coarse determination of their relative positions must be made before any Bragg patterns can be analyzed. Pseudo powder patterns were synthesized for this purpose by summing a large number of thermolysin diffraction images, all recorded at the same sample-detector distance. A graphical application was written, permitting the manual adjustment of the quadrant locations to align the observed powder rings with overlaid circular fiducial rings. This program is also suitable for calibrating the detector quadrants with silver behenate[53] powder patterns.

Prior to the experiment, the sensor positions and orientations (within each quadrant) were characterized optically at the LCLS to within tens of µm, but this calibration did not necessarily achieve the accuracy required for spot modeling, nor did it probe the actual readouts that are bump-bonded to the sensors. Each sensor is bonded to a pair of side-by-side 194 x 185 pixel application-specific integrated circuits (ASICs)[26]. Detailed positions and orientations of the 64 ASIC readouts were refined by non-linear least squares refinement of the target functional

$$f = \sum_{\substack{ASICs,\\ crystals,\\ spots}} (\boldsymbol{r}_{\mathrm{obs}} - \boldsymbol{r}_{\mathrm{calc}})^2 \, ,$$

where $\boldsymbol{r}_{\mathrm{obs}}$ is the observed detector position of the Bragg spot centroid determined with the program *Spotfinder*[28], $\boldsymbol{r}_{\mathrm{calc}}$ is the modeled position after indexing, and the sum is over all *Spotfinder* spots (on all images and ASICs) that correspond to modeled spots. Variable parameters in the refinement included the positions and rotations of all ASICs, the position of the direct beam and crystal-to-detector distance for each crystal shot, and the orientation and unit cell dimensions for each crystal. Correct performance of this algorithm was monitored by considering the refined placement of pairs of ASICs bonded to the same silicon sensor, which are thought to be exactly aligned by a mechanical guide piece during the manufacture process. These internal controls derived from the thermolysin data (**Fig. 2a**) show that the ASIC pairs are mutually aligned to an r.m.s. rotation of 0.016° and an r.m.s. displacement perpendicular to the long sensor axis of 0.074 pixels; we interpreted these values as the accuracy limits of our refinement method. The tolerances were similar for the lysozyme data, 0.030° and 0.072 pixels respectively. In addition, we found that on the particular detector used for thermolysin, the 32 sensors had an r.m.s. tilt of 0.17° in the plane of the detector, and that the separation between same-sensor ASIC pairs varied with an r.m.s. deviation of 0.21 pixels (**Fig. 2a**).

Refinement of the detector distance

We calibrated the absolute distance between crystal sample and imaging detector to an accuracy of about 1 mm. Fortunately the indexing algorithm and indeed the entire data processing pipeline is robust to this level of uncertainty, with small errors in the distance being absorbed by other modeled parameters (unit cell dimensions, wavelength). We determined the distance by grid search around an initial estimate: an entire run collected at a fixed distance was reprocessed several times with calibration offsets differing by 0.5 mm, which were then scored by counting the number of images successfully indexed (**Supplementary Figure 2**). Offsets of ±8 mm from the best value reduced the indexing rate by roughly a factor of 2.

An alternate distance calibration is possible by observing circular powder patterns from silver behenate as noted above, and the *cctbx.xfel* software can faciliate this analysis. Such a calibration might offer improved accuracy as it uses a recognized standard, however, as a practical matter given the time constraints of collecting data at LCLS, it was more efficient to simply use the thermolysin or lysozyme data itself to estimate the distance as shown in **Supplementary Figure 2**.

Structure solution

Merged structure factors were phased by molecular replacement using *Phaser*[54] within the *Phenix*[55] system. For thermolysin, the search model consisted of thermolysin (PDB code 2TLI[56]) from which all non-protein atoms were removed; for lysozyme the model was taken from PDB code 4ET8[10]. New models were built into the resulting maps using *phenix.autobuild*[57], and further refined using *phenix.refine*[58]**.** Refinement statistics are shown in **Supplementary Table 1**. The molecular clashscore (number of bad all-atom overlaps per thousand atoms) and Ramachandran stereochemical statistics were calculated with *MolProbity*[59].

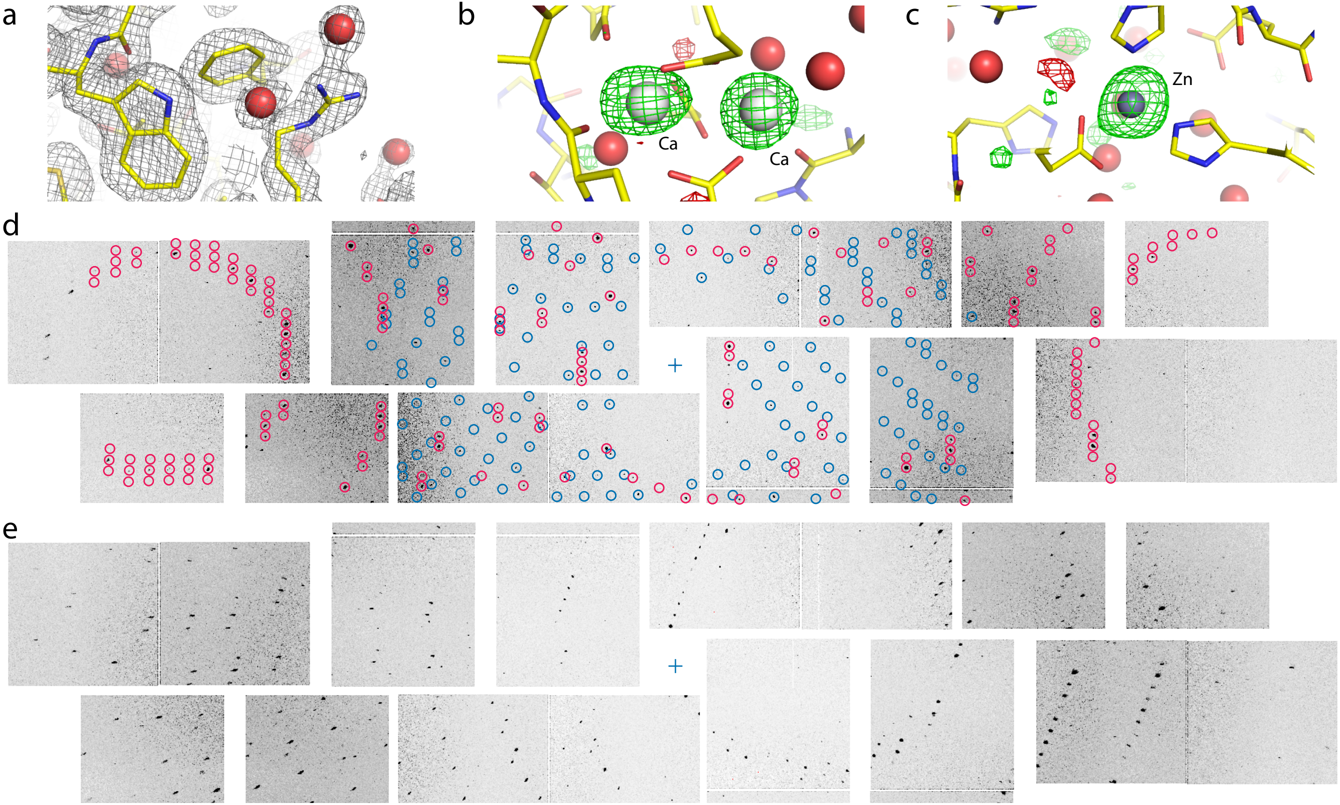
Crystallographic *R* factors for the refined thermolysin model are comparable in quality to synchrotron structures that have been determined at a similar resolution (2.1 Å). To determine this we used the program *phenix.r_factor_statistics*[60, 61] to print the *R* factor distribution from 2271 Protein Data Bank Structures at resolutions in the range 2.05–2.15 Å. Our thermolysin values of R_{work} = 22.2% and R_{free} = 26.5% are within one standard deviation of the mean (R_{work} = 20.1 ± 2.4%; R_{free} = 24.6 ± 2.6%). The *R* factor distribution was derived by taking coordinates, structure factors, and R-free flags from the Protein Data Bank, and using the *Phenix* toolbox to derive the *R* factors. As a result, the distributions can be directly compared with our refinements, which were also performed with *Phenix*.

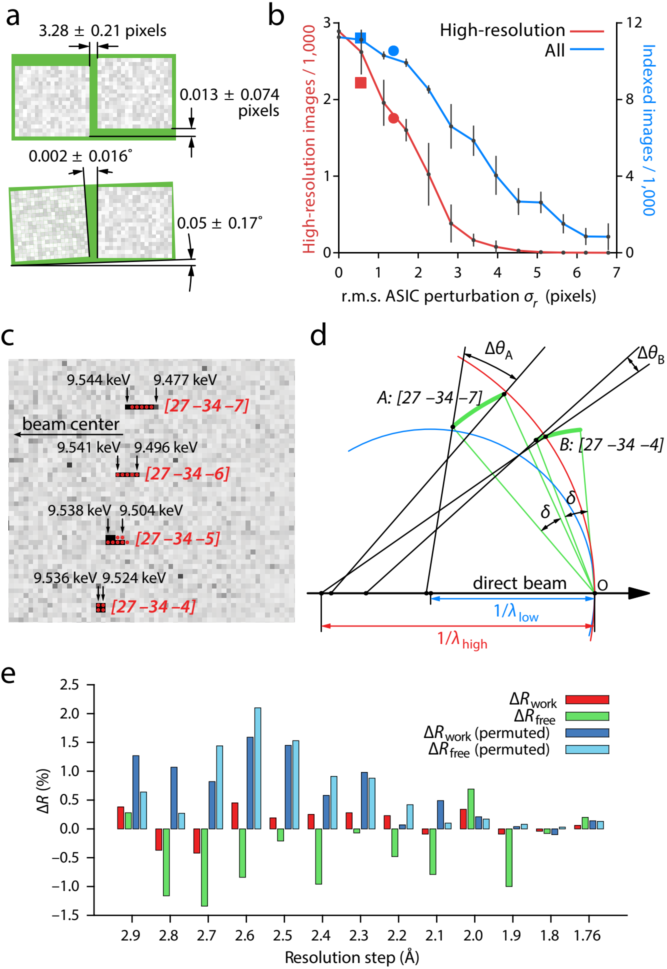
Similarly, for the 1.9 Å lysozyme structure, we considered 3578 Protein Data Bank Structures at resolutions in the range 1.85–1.95 Å. Our *Phenix*-refined values of $R_{\text{work}} = 18.7\%$ and $R_{\text{free}} = 22.9\%$ for the *cctbx.xfel* structure factors, and $R_{\text{work}} = 17.7\%$ and $R_{\text{free}} = 22.0\%$ for the *CrystFEL* structure factors, are each within one standard deviation of the mean ($R_{\text{work}} = 19.3 \pm 2.3\%$; $R_{\text{free}} = 23.2 \pm 2.6\%$).

The structure factors and model for thermolysin have been deposited with the Protein Data Bank under accession code 4OW3.

21. Inouye, K. *J. Biochem.* **112**, 335-340 (1992).
22. Titani, K. *et al. Nature* **238**, 35-37 (1972).
23. Boutet, S. & Williams, G.J. *New J. Phys.* **12**, 035024 (2010).
24. Sierra, R.G. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 1584-1587 (2012).
25. Bogan, M.J. *Anal. Chem.* **85**, 3464-3471 (2013).
26. Hart, P. *et al. Proc. of SPIE* **8504**, 85040C (2012).
27. Maia, F.R.N.C. *Nat. Methods* **9**, 854-855 (2012).
28. Zhang, Z. *et al. J. Appl. Crystallogr.* **39**, 112-119 (2006).
29. Steller, I., Bolotovsky, R. & Rossmann, M.G. *J. Appl. Crystallogr.* **30**, 1036-1040 (1997).
30. Rossmann, M.G. & van Beek, C.G. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1631-1640 (1999).
31. Sauter, N.K., Grosse-Kunstleve, R.W. & Adams, P.D. *J. Appl. Crystallogr.* **39**, 158-168 (2006).
32. Kern, J. *et al. Proc. Natl. Acad. Sci. USA* **109**, 9721-9726 (2012).
33. Giordano, R. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 649-658 (2012).
34. Diederichs, K. & Karplus, P.A. *Acta Crystallogr. D Biol. Crystallogr.* **69**, 1215-1222 (2013).
35. Paithankar, K.S. *et al. Acta Crystallogr. D Biol. Crystallogr.* **67**, 608-618 (2011).
36. White, T.A. *et al. Acta Crystallogr. D Biol. Crystallogr.* **69**, 1231-1240 (2013).
37. Nave, C. *Acta Crystallogr. D Biol. Crystallogr.* **54**, 848-853 (1998).
38. Otwinowski, Z. & Minor, W. *Methods Enzymol.* **276**, 307-326 (1997).
39. Emma, P. *et al. Nature Photon.* **4**, 641-647 (2010).
40. Greenhough, T.J. & Helliwell, J.R. *J. Appl. Crystallogr.* **15**, 338-351 (1982).
41. Greenhough, T.J. & Helliwell, J.R. *J. Appl. Crystallogr.* **15**, 493-508 (1982).
42. Greenhough, T.J., Helliwell, J.R. & Rule, S.A. *J. Appl. Crystallogr.* **16**, 242-250 (1983).
43. Ren, Z. & Moffat, K. *J. Appl. Crystallogr.* **28**, 461-481 (1995).
44. Dauter, Z. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 1703-1717 (1999).

45. Diederichs, K. *Acta Crystallogr. D Biol. Crystallogr.* **65**, 535-542 (2009).
46. Schreurs, A.M.M., Xian, X. & Kroon-Batenburg, L.M.J. *J. Appl. Crystallogr.* **43**, 70-82 (2009).
47. Porta, J. *et al. Acta Crystallogr. D Biol. Crystallogr.* **67**, 628-638 (2011).
48. Bolotovsky, R. & Coppens, P. *J. Appl. Crystallogr.* **30**, 65-70 (1997).
49. Leslie, A.G.W. *Acta Crystallogr. D Biol. Crystallogr.* **62**, 48-57 (2006).
50. Kahn, R. *et al. J. Appl. Crystallogr.* **15**, 330-337 (1982).
51. Brünger, A.T. *Nature* **355**, 472-475 (1992).
52. Strüder, L. *et al. Nucl. Instrum. Methods Phys. Res. A* **614**, 483-496 (2010).
53. Huang, T.C. *et al. J. Appl. Crystallogr.* **26**, 180-184 (1993).
54. McCoy, A.J. *et al. J. Appl. Crystallogr.* **40**, 658-674 (2007).
55. Adams, P.D. *et al. Acta Crystallogr D* **66**, 213-221 (2010).
56. English, A.C. *et al. Proteins: Structure, Function, and Genetics* **37**, 628-640 (1999).
57. Terwilliger, T.C. *et al. Acta Crystallogr. D Biol. Crystallogr.* **64**, 61-69 (2008).
58. Afonine, P.V. *et al. Acta Crystallogr. D Biol. Crystallogr.* **68**, 352-367 (2012).
59. Chen, V.B. *et al. Acta Crystallogr. D Biol. Crystallogr.* **66**, 12-21 (2010).
60. Urzhumtseva, L. *et al. Acta Crystallogr. D Biol. Crystallogr.* **65**, 297-300 (2009).
61. Afonine, P.V. *et al. J. Appl. Crystallogr.* **43**, 669-676 (2010).

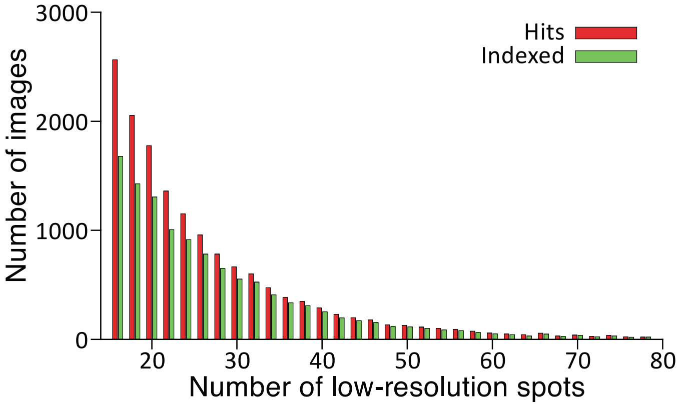

a
b
Ca
Ca
c
Zn
d
+
e
+


a
3.28 ± 0.21 pixels
0.013 ± 0.074 pixels
0.002 ± 0.016˚
0.05 ± 0.17˚
b
High-resolution
All
High-resolution images / 1,000
Indexed images / 1,000
0
1
2
3
0
4
8
12
0
1
2
3
4
5
6
7
r.m.s. ASIC perturbation σ_r (pixels)
c
9.544 keV
9.477 keV
[27 −34 −7]
beam center
9.541 keV
9.496 keV
[27 −34 −6]
9.538 keV
9.504 keV
[27 −34 −5]
9.536 keV
9.524 keV
[27 −34 −4]
d
$\Delta\theta_A$
$\Delta\theta_B$
A: [27 −34 −7]
B: [27 −34 −4]
δ
δ
direct beam
O
$1/\lambda_{low}$
$1/\lambda_{high}$
e
ΔR_{work}
ΔR_{free}
ΔR_{work} (permuted)
ΔR_{free} (permuted)
ΔR (%)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
2.9
2.8
2.7
2.6
2.5
2.4
2.3
2.2
2.1
2.0
1.9
1.8
1.76
Resolution step (Å)

Accurate macromolecular structures using minimal measurements from X-ray free-electron lasers

Johan Hattne, Nathaniel Echols, Rosalie Tran, Jan Kern, Richard J. Gildea, Aaron S. Brewster, Roberto Alonso-Mori, Carina Glöckner, Julia Hellmich, Hartawan Laksmono, Raymond G. Sierra, Benedikt Lassalle-Kaiser, Alyssa Lampe, Guangye Han, Sheraz Gul, Dörte DiFiore, Despina Milathianaki, Alan R. Fry, Alan Miahnahri, William E. White, Donald W. Schafer, M. Marvin Seibert, Jason E. Koglin, Dimosthenis Sokaras, Tsu-Chien Weng, Jonas Sellberg, Matthew J. Latimer, Pieter Glatzel, Petrus H. Zwart, Ralf W. Grosse-Kunstleve, Michael J. Bogan, Marc Messerschmidt, Garth J. Williams, Sébastien Boutet, Johannes Messinger, Athina Zouni, Junko Yano, Uwe Bergmann, Vittal K. Yachandra, Paul D. Adams, Nicholas K. Sauter

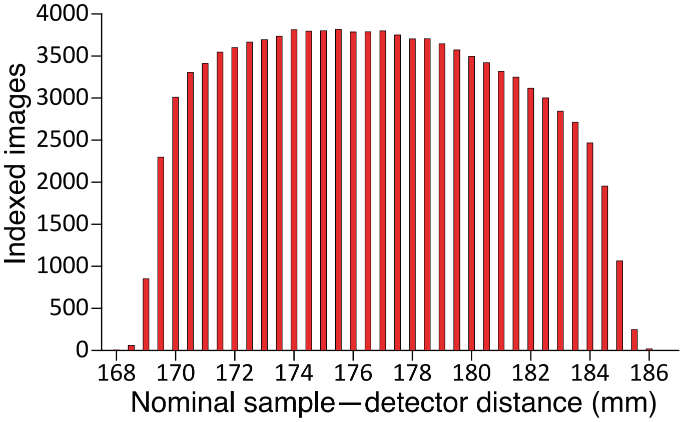
Supplementary Figure 1 | Indexing success.



Interval	Indexed/Hits	[%]
16-17	1674 / 2561	[65]
18-19	1423 / 2051	[70]
20-21	1302 / 1773	[73]
22-23	1003 / 1357	[74]
24-25	911 / 1148	[79]
26-27	778 / 955	[82]
28-29	646 / 778	[83]
30-31	549 / 662	[83]
32-33	522 / 597	[87]
34-35	405 / 470	[86]
36-37	332 / 381	[87]
38-39	305 / 344	[89]
40-41	250 / 286	[87]
42-43	194 / 226	[86]
44-45	168 / 195	[86]
46-47	151 / 175	[86]
48-49	114 / 130	[88]
50-51	112 / 126	[65]
52-53	98 / 109	[70]
54-55	84 / 96	[73]
56-57	77 / 88	[74]
58-59	59 / 72	[79]
60-61	48 / 55	[82]
62-63	40 / 47	[83]
64-65	28 / 39	[83]
66-67	45 / 53	[87]
68-69	22 / 28	[86]
70-71	33 / 36	[87]
72-73	20 / 23	[89]
74-75	28 / 33	[87]
76-77	16 / 20	[86]
78-79	18 / 18	[86]
All	11455/14932	[77]

Within the thermolysin dataset, 14932 diffraction patterns were identified as having between 16 and 79 candidate Bragg spots at low angles out to 4.0 Å resolution. Of these, 11455 were ultimately indexed and integrated.

Supplementary Figure 2 | Distance calibration.



Distance (mm)	Indexed
168.0	5
168.5	60
169.0	852
169.5	2297
170.0	3010
170.5	3304
171.0	3411
171.5	3546
172.0	3601
172.5	3665
173.0	3695
173.5	3735
174.0	3813
174.5	3796
175.0	3800
175.5	3817
176.0	3785
176.5	3788
177.0	3798
177.5	3749
178.0	3704
178.5	3705
179.0	3645
179.5	3571
180.0	3495
180.5	3420
181.0	3315
181.5	3248
182.0	3117
182.5	3000
183.0	2845
183.5	2713
184.0	2466
184.5	1954
185.0	1064
185.5	248
186.0	18
186.5	1

Thermolysin data from run 21 were reprocessed with different trial distances, with the final value of 175.5 mm chosen on the basis of indexing success rate. After this calibration, a beamline encoder provided relative offsets for data collected at different detector distances.

Supplementary Table 1 | Data collection and refinement statistics.

	Thermolysin processed with *cctbx.xfel*	**Lysozyme, 40fs exposure processed with *cctbx.xfel***	**Lysozyme, 40 fs exposure processed with *CrystFEL***
Data collection			*as reported in ref. 12:*
Mean wavelength (Å)	1.269 ± 0.001 (N = 12,692) 1.297 ± 0.001 (N = 912)	1.320 ± 0.002 (N = 21,743)	1.32
Space group	$P6_122$	$P4_32_12$	$P4_32_12$
Cell dimensions[a]			
a, c (Å)	92.9 ± 0.3, 130.4 ± 0.6	79, 38	79, 38
Resolution[b] (Å)	68.5–2.10 (2.18–2.10)	39.5–1.90 (1.97–1.90)	35.3–1.90
No. collected images	651,793	1,507,834 (runs 305-327)[c]	1,471,615 (runs 305-327)
No. images used	11,647	21,743	12,247
No. lattices merged	13,371	23,929	12,247
No. total reflections	19,854 (1,781)	9923 (948)	9921
R_{split}[d](%)	24.4 (79.2)	13.0 (25.0)	15.8
$CC_{1/2}$[e](%)	84.5 (17.0)	96.0 (75.2)	n.a.
CC_{iso}[f](%)	85.5 (36.9)	82.9 (84.3)	n.a.
$I / \sigma (I)$[g]	50.2 (5.6)	97.1 (22.8)	7.4 (2.8)
Completeness (%)	99.1 (91.2)	100.0 (100.0)	98.3 (96.6)
Multiplicity	209.0 (4.5)	587.0 (125.7)	n.a.
Wilson B factor (Å^2)	14.3	14.1	28.3
Refinement[h]			*as re-refined by us:*
R_{work} / R_{free} (%)	22.2 / 26.5 (32.2 / 37.4)	18.7 / 22.9 (19.2 / 26.2)	17.7 / 22.0 (20.2 / 33.7)
No. atoms			
Protein	5094	1001	1001
Ligand/ion	5	2	2
Water	452	137	84
B-factors (Å^2)			
Protein	14.7	13.7	30.6
Ligand/ion	14.6	18.3	38.0
Water	24.1	28.1	40.2
R.m.s. deviations			
Bond lengths (Å)	0.003	0.004	0.008
Bond angles (°)	0.69	1.04	1.25
Clashscore[i]	0.86	3.06	3.06
Ramachandran statistics			
Favored (%)	96	98	99
Outliers (%)	0	0	0

[a] For thermolysin the distribution of unit cell dimensions was experimentally determined, leading to the population standard deviation given here. Unit cell dimensions for lysozyme were freely determined during indexing but constrained during merging to the same values reported in ref. 12, to facilitate a comparison between *cctbx.xfel* and *CrystFEL*.

[b] The high-resolution cutoff for thermolysin was determined as described in the text. For lysozyme it was constrained to the value (1.9 Å) reported in ref. 12 to facilitate the comparison. Statistics reported in parentheses represent values computed for the highest resolution shells (see **Supplementary Tables 2 and 3**).

[c] The run numbers represent the raw-data file serial numbers we believe were actually used to derive the ref. 12 structure factors, based on the diffraction image list deposited in the Coherent X-ray Imaging Data Bank (http://www.cxidb.org/data/17/40fs_5fs_indexed.txt) and the number of images (12,247) reported in the ref. 12 paper. The number of collected images we report (1,507,834) is the number actually present in those data files, which differs slightly from that (1,471,615) listed in the paper.

[d] R_{split} measures the percent difference between half-datasets as defined in ref. 12.

[e] $CC_{1/2}$ is the correlation coefficient between half-datasets defined in ref. 21.

[f] CC_{iso} is the correlation coefficient with a reference set of structure factor intensities. See **Supplementary Tables 2 and 3** for details.

[g] $I / \sigma (I)$ values from the *cctbx.xfel* and *CrystFEL* programs can not be directly compared; see the **Online Methods**.

[h] In the lysozyme/*CrystFEL* column the refinement statistics represent our re-processing of the ref. 12 structure factors (as deposited in Protein Data Bank entry 4ET8) using our protocols, in order to control for any differences between our structure solution procedures and those employed by the other group. For both lysozyme refinements we used the same R_{free} flags as were originally used in ref. 12.

[i] Clashscore is the number of bad all-atom clashes per thousand atoms from *MolProbity*[55].

Supplementary Table 2 | Thermolysin merging statistics by resolution bin.[a]

Resolution	#	#	# Unique	Complete-	<Multiplicity>	<*I*/σ(*I*)>	R_{split}(%)	$CC_{1/2}$	CC_{iso}[c]
range (Å)	Lattices[b]	Measurements	reflections	ness (%)				(%)	(%)
∞ – 4.53	13,371	1,589,278	2,196	100.0	723.7	123.4	15.4	90.6	86.0
4.5 – 3.59	11,678	1,008,350	2,050	100.0	491.9	123.8	16.2	86.7	89.8
3.59 – 3.14	10,300	627,340	2,014	100.0	311.5	78.7	18.1	85.6	88.5
3.14 – 2.85	8,326	418,197	1,998	100.0	209.3	55.4	19.3	82.9	86.6
2.85 – 2.65	6,770	249,160	1,978	100.0	126.0	36.9	23.8	79.8	86.6
2.65 – 2.49	5,482	130,094	1,983	100.0	65.6	24.2	29.7	70.0	83.0
2.49 – 2.37	4,416	66,763	1,955	100.0	34.2	16.6	39.9	55.5	76.6
2.37 – 2.26	3,556	34,321	1,953	100.0	17.6	11.8	52.6	27.7	64.0
2.26 – 2.18	2,648	17,632	1,946	99.2	9.1	8.3	62.7	36.1	59.9
2.18 – 2.10	1,700	8,009	1,781	91.2	4.5	5.6	79.2	17.0	36.9
∞ – 2.10	13,371	4,149,144	19,854	99.1	209.0	50.2	24.2	84.5	85.5
2.10–2.03[d]	529	3,300	1,246	63.8	2.7	4.3	86.6	14.5	23.2
2.03 – 1.98	225	1,935	910	47.4	2.1	3.6	90.3	53.9	28.9
1.98 – 1.92	209	1,241	696	35.4	1.8	3.0	91.4	2.8	27.6
1.92 – 1.88	191	740	484	25.1	1.5	2.8	85.4	41.9	n.a.
1.88 – 1.84	149	409	303	15.6	1.4	1.8	99.8	6.5	n.a.
1.84 – 1.80	23	59	59	3.1	1.0	1.5	n.a.	n.a.	n.a.
1.80 – 1.76	9	14	13	0.7	1.1	1.9	n.a.	n.a.	n.a.
∞ – 1.76	13,371	4,156,853	23,565	70.1	176.4	42.8	26.0	79.2	81.1

[a] In **Supplementary Tables 2 and 3**, resolution bins are chosen so as to have equal reciprocal space volumes, which is the default option in the *cctbx* toolkit.

[b] This column reports the number of lattices that have been successfully indexed and integrated and that contain any integrated observations in the indicated resolution range. The full set of 13,371 lattices was used to produce the merged thermolysin data in **Supplementary Table 1**. Successively smaller nested subsets of this collection contained data in successively higher resolution bins, with the smallest nested subset (of only 1700 lattices) extending to 2.1 Å. We refer to "lattices" instead of "images" because some images contain diffraction lattices from two distinct microcrystals (**Fig. 1d**) that can be usefully integrated with *cctbx.xfel*.

[c] Correlation coefficient between the XFEL-observed structure factor intensities and the previously published structure factors from Protein Data Bank entry 2TLI. The α–carbon r.m.s.d. between the two respective models is 0.41 Å. Both structures were determined at room temperature.

[d] Data in the 2.10 – 1.76 Å resolution ranges were not included in the final averaged structure factor set, because they do not satisfy our inclusion criterion (**Fig. 2e**). They are listed here simply to document how many images were observed to contain high-resolution spots in the detector corners. Note that the scarce data in these resolution ranges have insufficient multiplicities (1.0 – 2.7) for accurate averaging, which is also reflected in the poorly-behaved $CC_{1/2}$ and R_{split} statistics that compare merged half-datasets.

Supplementary Table 3 | 40 fs-pulse lysozyme merging statistics by resolution bin, for the *cctbx.xfel*-processed data.[a]

Resolution range (Å)	# Lattices[b]	# Measurements	# Unique reflections	Completeness (%)	<Multiplicity>	<$I/\sigma(I)$>	R_{split}(%)	$CC_{1/2}$ (%)	CC_{iso}[c] (%)
∞ – 4.09	23,929	1,443,414	1,089	99.9	1325.5	215.5	10.8	95.8	78.8
4.09 – 3.25	23,260	1,153,093	1,022	100.0	1128.3	208.4	10.5	94.0	95.0
3.25 – 2.84	22,014	838,213	999	100.0	839.1	137.3	12.3	92.9	95.8
2.84 – 2.58	19,768	584,099	989	100.0	590.6	96.5	13.4	91.7	94.3
2.58 – 2.39	17,230	529,545	978	100.0	541.5	82.0	12.9	91.4	93.6
2.39 – 2.25	15,513	449,561	977	100.0	460.1	67.6	15.8	88.1	89.1
2.25 – 2.14	13,672	314,469	967	100.0	325.2	49.5	17.1	89.3	90.7
2.14 – 2.05	11,486	231,022	986	100.0	234.3	39.1	18.7	86.8	90.6
2.05 – 1.97	9,599	161,713	968	100.0	167.1	29.7	21.5	81.7	89.6
1.97 – 1.90	7,925	119,195	948	100.0	125.7	22.8	25.0	75.2	84.3
∞ – 1.90	23,929	5,824,324	9,923	100.0	587.0	97.1	13.0	96.0	82.9

[a] In **Supplementary Tables 2 and 3**, resolution bins are chosen so as to have equal reciprocal space volumes, which is the default option in the *cctbx* toolkit.

[b] The full dataset contained 21,743 images successfully indexed and integrated. Some contained diffraction lattices from two distinct microcrystals, thus giving a total of 23,929 lattices.

[c] Correlation coefficient between the *cctbx.xfel*-processed structure factor intensities and the published *CrystFEL*-processed structure factors from Protein Data Bank entry 4ET8. The α–carbon r.m.s.d. between the two respective models is 0.08 Å.

Supplementary Note | Targeting the exact pixels that contain signal

The concept of tailoring the Bragg spot shape model to the data at hand is highlighted by the recent work of Kirian *et al.*[14] that analyzes photosystem I (PS I) XFEL XRD[4]. The observed diffraction fringes connecting each low-angle Bragg spot with its six nearest neighbors arise when the crystallite contains only a small number of unit cells along each crystal axis. Therefore a main challenge in that work was to select just the central portion of each Bragg peak for data integration. The width of the central peak scales as λ/a (where λ is the wavelength of the incident light, and a is the width of the crystal). Accordingly, the *CrystFEL* program was developed with a single adjustable parameter, $\delta \approx 1/a$, and it considers a pixel to contain signal if the lattice model places the corresponding reciprocal point within reciprocal distance δ of the sphere of reflection (the Ewald sphere). A single δ radius is chosen that best models the entire ensemble of diffraction images. The treatment of Bragg spots as reciprocal space spheres instead of complicated fringe functions is an approximation well-suited to the 8.5 Å resolution PS I data[4]. In contrast, the high-resolution experiments described here use hard X-rays (λ ~ 1.3 Å, instead of 7 Å), and crystal sizes are larger by a factor of ~10, reducing the λ/a broadening to less than one pixel. The size and shape of Bragg spots is determined by other effects, familiar from SR crystallography, including the presence of microscopic "mosaic" crystal domains, which cause Bragg spots to appear as concentric arcs if there is a distribution of orientations, or as large circles if there is a distribution of cell dimensions. Also, XFEL experiments performed with self-amplified spontaneous emission (SASE) pulses have a considerably larger bandwidth than SR experiments, causing Bragg spots to be streaked in the radial direction. The result is that Bragg spots in high-resolution XFEL experiments are markedly different from the integration masks of constant radius currently used by *CrystFEL,* particularly at higher diffraction angles.